Engineering Language Resources and Technologies

Point of Contact: Damir Cavar’ (dcavar@indiana.edu), Department of Linguistics, College of Arts and Sciences, IUB, Memorial Hall, phone: 812 391-3602 or 734 709-6039

The Grand Challenge is to engineer new Universal Speech Processing Technologies and Video-Analysis Algorithms capable of processing spoken human language in natural environments without language-specific or domain-dependent models.

Abstract

We propose focusing research on universal technologies for speech analysis that will significantly facilitate the engineering of speech and language technologies (SLTs) for the remaining 99% of languages for which common speech and language technologies do not exist. This technology will be accompanied by processing of visual information, i.e. semantic analysis of situation and event contexts.

SLTs face serious limitations due to their dependence on large amounts of language-specific resources that are costly to produce and largely unavailable. Also current SLTs have been engineered using a relatively narrow footprint of languages, being restricted to well-resourced languages. They also do not include visual information available in natural dialogue scenarios, for example in video recordings, to exploit visual cues in the environment, or even gesture and body language, which are useful cues for processing natural language. There is a growing demand for wider cross-linguistic SLT coverage. The expansion of audio and video data in new communication technologies throughout the world demands solutions that are more language-independent and scalable.

IU is ideal for this project given its unusual combination of strengths in language-related research and education, computer science, informatics, and future-oriented domains like cybersecurity and engineering.

The goals are to gain insights in universal linguistic properties relevant for SLTs that can be integrated into a new generation of algorithms, and to demonstrate how new types of SLT can interact with visual processing in real human conversation. The results will be demonstrated in applications geared towards medical or emergency services for under-resourced languages.

The Grand Challenge

There is no universal, language-independent technology for mapping spoken language and accompanying gesture, body language, and setting cues onto a representation that can be computationally analyzed using Natural Language Processing technologies. The majority of

1 If this GC is selected for funding, Damir Cavar would in all likelihood be a tenured Associate Professor and hence eligible, and in the event that he would not, another tenured faculty member would take his place, e.g. professors Sandra Kuebler, Markus Dickinson, or Kenneth de Jong from the Departments of Linguistics and Cognitive Science.
language technology research is concerned with text processing, while written text is just a reflection of spoken language and many languages of the world do not have a writing system. We are proposing a paradigm shift towards human-centric processing of language with a focus on speech and gesture. By making this shift, we refocus attention to general aspects of human language function which are the foundation for all spoken languages. Current speech and language technologies (SLTs) depend on large amounts of language-specific data in appropriate formats with extensive annotations. Such data is available for less than 1% of all languages and it is time consuming and expensive to produce. This is the so-called “language resource bottleneck” in SLT research and engineering. As digital audio and video data grows exponentially, it encompasses a growing number of languages from all around the world, most of which do not have appropriate resources to build SLTs using current methods. To handle this expansion, we need to find efficient algorithms for SLTs and language-related visual information that are language-independent and computationally scalable. These technologies are essential for breaking through the language resource bottleneck and opening up SLT development for all of human languages. They are essential core technologies that will play a major role in the competitiveness of IU in the high-tech, international, and interdisciplinary sector of speech and language-enabled technology research. Globally, the resources and technologies represent an invaluable asset for IU as they are of high importance for national and cyber security, as well as for education, language and linguistic research, and even for the development of civil societies throughout the globe.
Goals

All SLTs are based on language models that are extracted from natural language resources such as speech and text corpora, augmented with meta-information and annotated for linguistic properties. These resources are used to generate quantitative and qualitative models for various SLTs. The first goal is to **build an open-ended infrastructure to collect and store digital spoken and textual language corpora (see Team 1 and 2)** (audio, video, transcriptions, annotations) that are useful for SLT research and engineering, the generation of material for language education, and language-related research. This infrastructure is intended to become a sustainable and highly valuable service at IU. The Archive of Traditional Music (ATM), Library Technologies, and IT-Services already offer core services that can be complemented with the proposed service for resource management for language-related research and SLT engineering.

The second goal is to establish the missing infrastructure for SLT research and engineering at IU in the form of a **Speech and Language Technology Institute (see Team 3 and 4)**. Its role will be to connect the corpora generated as the first goal to the technological applications by managing the collection and generation of speech and language resources (digital data and corpora) that are designed for researching and engineering SLTs. It will provide an environment for experiments with speech processing technologies, algorithms, and models. With a focus on managing valuable assets of research and technology engineering data and a research focus on algorithms and machine learning technologies geared for SLT and Human Language Technology (HLT) applications, it is intended to become a sustainable institution at IU with the ability to attract external funding and private sector cooperation.

Initial experiments with generic and universal speech recognizers will be based on corpora provided by partnering language departments and centers currently existing at IU. The corpora consist of time-aligned transcription at the word and sentence level. They will be used to train **Forced Alignment technologies (see Team 4)** to generate signal synchronized human-action based transcription automatically. This will facilitate the generation of speech corpora for training **Automatic Speech Recognition (ASR) systems (also see Team 4)**. During this phase we will perform experiments with the extraction of phonotactic units (i.e. specific sound production sequences) that have a wide cross-linguistic distribution. At the same time, the video material will serve initial **experiments with scene and event recognition and labeling (see Team 5)**. In both sets of experiments the focus will be generating models for ASRs and video taggers based on various types of feature extraction from the speech signal or video stream.

The resulting algorithms and SLTs will be tested using applications for the areas outside SLT, Artificial Intelligence, or linguistic research. We will demonstrate how the generated resources can be used for language education, using not only extracted data, but also SLTs and HLTs geared at evaluation and learning processes. We will demonstrate how the technologies can be applied to serve Digital Humanities, Library Technologies, and Archives in the **automatic analysis and annotation of digital audio and video resources** that contain speech and video from a variety of languages. Additionally we will demonstrate uses of **SLTs in the domain of medical consultation and emergency call dispatching**.

**Proposed research and its impact**
This proposal emerged from ongoing research activities and various previous research projects funded for example by NSF. In particular the NSF-supported grant “Automatically Annotated Repository of Digital Audio and Video Resources Community” (#1244713) contributed to the current proposal and defined research strategies. During this exploratory and community building grant we had the opportunity to evaluate current SLTs and video processing technologies and their properties and shortcomings, in particular, with respect to audio and video processing with speech content from under-resourced languages. In four interdisciplinary workshops we discussed the current research issues and future directions for speech, language and video analysis technologies with international experts from informatics, computer science, speech and language technologies, and humanities, as well as archivists and library technology and service providers.

The proposed research project on universal speech and language, and video tagging technologies for the purpose of speech recognition and analysis for all languages is targeting various serious research issues that are related to the earlier mentioned a) language resource bottleneck, as well as b) dynamic properties of languages, c) the need for language-independent algorithms, technologies and language models that are potentially based on language universals or insights about cognitive processes related to perception of spoken language or visual information, d) computational semantics algorithms and technologies that address the problems of efficient reasoning over large knowledge representations, e) the role of interacting technologies that process parallel independent information streams like speech and video exchanging information for content analysis and disambiguation, using potentially adaptive machine learning solutions, and f) the role of efficient algorithms, measured in terms of resources like CPU-operations and memory demand, for use on “big audio and video data”.

Our research and solutions for the language resource bottleneck problem will include:

1. new technologies or strategies to bootstrap language resources with annotations (corpora, acoustic and language models) for under-resourced languages by using resources from related or similar well-documented languages;
2. adaptation of existing technologies and algorithms to become less language- and content-specific using language universals or linguistic knowledge, insights from cognitive and phonetic research, and machine learning strategies;
3. remodeling the speech and language processing strategies to involve independent recognition phases for acoustic speech properties (a level of phonotactic units) and subsequent mapping on language properties and models (words, phrases and orthographic representations);
4. integration of knowledge representations, deep reasoning, and general visual cue processing in the speech and language processing chain.

Common speech recognition systems utilize probabilistic models of the distribution of acoustic properties of the speech signal associated with symbolic representations of sounds. To improve the accuracy of the recognition, we will use language models to filter the most likely hypotheses about underlying sound sequences. We will extend this interaction of acoustic and language models with a component that involves visual cues for disambiguation. In addition, we will experiment with the inclusion of universal ontological knowledge representations and reasoning algorithms, and online visual cue extraction and processing in the speech and language processing algorithms to improve recognition rates and linguistic processing precision.
The sub-tasks and responsibilities are listed in the following:

- **Team 1 and 2: Data collection and preparation.** This part of the project includes conversion of collected language data to interoperable and machine processable corpora. This includes: a) development and support of digital repository systems for preservation of and access to language data; b) building on IU’s existing services and R&D in these areas, including the Avalon Media System (avalonmediasystem.org), IUScholarWorks (scholarworks.iu.edu), and IU’s partnership in the Hydra community (projecthydra.org); c) facilitating access to MDPI audio and video data; d) integrating tools and technologies developed by the project with local workflows and systems to improve access to MDPI and other spoken language-based audiovisual content held by IU; e) further into the process, the Libraries will also play a role in education and outreach to communicate research outcomes; and f) the resources will be linked with international projects in Australia, Europe and Asia.

- Generating language models and acoustic models. Collected corpora in a usable format will serve to generate language models and acoustic models. These will be further used for individual language models but also to experiment on optimization of algorithms and development of universal algorithms.

- Optimizing algorithms for ASR. The idea is to detect strategies how to use already developed language models and acoustic models, how to fine-tune them to be used on languages which are structurally similar, and how to incorporate linguistic knowledge to optimize algorithms.

- Video recognition algorithms. The goal here is to develop and optimize video recognition algorithms, and to research strategies for the use of video annotation for the interpretation of the accompanying spoken word.

- Developing applications of the developed language materials for the use in education, translation.

- Test case: development of transcriptions, translations and/or annotations of audio and video collections in ATM.

**Practical mechanisms to ensure impact in the real world.** The creation of the dedicated administrative unit, a Speech and Language Technology Institute, is central to the project in order to service the administration of the projects and facilitate the coordination between the sub-projects. As an offshoot of the project activities, a huge repository of language resources will be created, that will hold resources in a unified format allowing for cross-language research/approaches for future projects. Such a repository (GORILLA initiative, gorilla.linguistlist.org) has already been initiated and will be soon populated with few pilot resources that have been created by the faculty affiliated with the LINGUIST List during the summer of 2015. Researchers associated with some industries were interested in such resources and a partnership on the development of applications.

**Impact.** We expect the project to have a broad impact locally for the Bloomington campus and IU, generally for the state of Indiana, as well as globally.

**Impact for IU and the Bloomington campus:**

- The project will produce new synergies resulting from the multidisciplinary approach spanning across Arts and Sciences (Humanities, Linguistics, and individual language
departments), Computer Science and Informatics, with possibility to serve Medicine, Law, and Business, and positively affect research activities at IU.

- The development of language technology of direct application to communities worldwide will increase Indiana University’s profile in developing countries, such as India and China, which are moving toward the forefront of technological development. This will increase productive connections to computational and language research in these developing countries, as well as interest among the most promising young scholars in these countries for studying at Indiana University.
- The project links STEM technologies with the Humanities, creating new research opportunities, increasing the competitiveness for external funding.
- It creates links between university research and the private sector, potentially enabling co-funding or collaborations in terms of joint grants, skill transfer, or internships for students.
- It is using and building on the unique resources and the strengths of IU, specifically, the current, unparalleled cohort of language and culture resources from many under-resourced language groups.
- It creates new education opportunities and consequently new career paths for students.
- New assets are created at IU of high international and industrial value, including language resources, speech and language technologies, language corpus tools and other techniques.
- The materials created by the Digitization and Preservation initiative at IU are safe and preserved but still not available for large scale research and analysis. The current project will enable the created materials for automatic search and analysis.

**For the state of Indiana:**

- Unique resources are created locally that can be used by Indiana residents, students, and entrepreneurs.
- The project has high potential for investments and private sector partnerships with SLT industry (e.g. ...), and a high potential for new patents start-up spin-offs and job creation.
- It will create resources and tools necessary for the development of applications that can improve life at various levels, among others, applications using speech technologies in medical, security, emergency sector, language technologies for machine translation, information extraction, and cyber-security.
- It creates resources and technologies for languages present in the State of Indiana, where many of these languages are not documented, and technological means or resources do not exist.

**Globally:**

- The project will result in extending language technologies to new languages and new communities worldwide. This will improve our ability to provide resources and technologies in emerging incidents in different regions of the world (technologies and language data for domains like medicine, natural disasters, security, and military purposes). Access to speech and language technologies will enable minority communities’ participation in technological progress and strengthen civil society initiatives worldwide.
A large amount of so-far untranscribed, non-annotated and non-translated audio and video resources will become available for large-scale search and research. This includes digital recordings in archives, historical, cultural documents available through libraries but also media on internet which are interesting for linguists, historians, sociologists, business, and medical personnel.

Resources
This proposal builds on and spans across unique strengths of the Bloomington campus: a large number of language experts and language programs built at great expense and persistence over the last 50 years, a strong focus on informatics and computing, and language resources in various departments and centers. It is also in line with the newest IU initiative regarding cybersecurity, Media Digitization and Preservation Initiative, and the envisioned engineering program at the Bloomington campus. Given this, IU is an ideal environment for interdisciplinary research on SLT using the synergies between existing units of the Bloomington campus.

Language experts. The Bloomington campus hosts numerous centers and institutes dedicated to the research on individual languages and cultures or focused on language groups or geographic areas, and grouping experts on particular languages. More than 70 languages are taught on a regular basis on the campus. Many of these centers develop their own teaching materials and can serve as an easily available on-site source of expertise and language data such as recordings and texts, which can supplement the materials provided by the communities and researchers worldwide. Additionally, IU hosts approximately 7,000-8,500 international students and scholars each year, many of them speakers of under-resourced languages. This means that we have locally invaluable resources for specific languages: native speakers and language experts knowledgeable about rare languages.

Language resources. IU hosts also established archives, some of them curating language resources, and among them is the Archive of Traditional Music (ATM). Further, within Media Preservation and Digitization Initiative, valuable resources which had been seriously endangered due to media degradation and format obsolescence, many of them in languages other than English, are being digitized. The digitized resources remain safe but are still unusable for large scale research without any accompanying transcriptions. These transcriptions could be delivered using automatic SLT systems, opening previously inaccessible resources for search and research.

Information science/informatics experts/computational linguistics. IU is home to School of Informatics and Computing, one of the finest such institutions in the country, with experts in informatics, information science, and intelligent systems engineering Linguistics Department of CAS holds well-established programs in computational linguistics.

The Missing Link/Strategic Hires is support for speech engineering through a number of faculty lines, postgraduate researchers and graduate student support. The Bloomington campus is also missing experts for computational semantics, knowledge representation and processing. There is a growing need for experts in processing, analysis or summarization of visual information.
Team
Point of contact: Damir Cavar <dcavar@indiana.edu>. The suggested team leaders are marked in bold.

Team 1: Libraries and archiving
Jon Dunn (Interim Assistant Dean of Library Technologies), Nick Homenda (Assistant Librarian), Karen Farrell (Librarian for South Asian and Southeast Asian Studies), Alan Burdette (Adjunct Professor and Director of Archive of Traditional Music)

Team 2: Language data, corpora use for research in humanities
John McDowell (Professor and Chair of Folklore and Ethnomusicology), Sarah Phillips (Professor of Anthropology and Director of REEI), Mark Trotter (Associate Director of REEI), Charles Lin (Assistant Professor of Chinese Linguistics in East Asian Languages & Cultures), Dov-Ber Kerler (Professor and Chair in Yiddish Studies), Fritz Breithaupt (Professor and Chair in Germanic), Oner Ozcelik (Director and Assistant Professor in central Eurasian Studies, CeLCAR), Clara Henderson (Director of Institute for Digital Arts and Humanities), Julie Auger (Associate Professor in French and Linguistics), Malgorzata Cavar (Assistant Professor of Linguistics and Co-Director of the LINGUIST List), Manuel Diaz-Campos (Professor of Spanish)

Team 3: Research on universal properties of languages that can be used in the development of SLT tools
Damir Cavar (Associate Research Scientist and Co-Director of the LINGUIST List), Malgorzata Cavar (Assistant Prof.), Kelly Berkson (Assistant Professor), Ken de Jong (Professor)

Team 4: Research on ASR algorithms
Secondary: Damir Cavar, Malgorzata Cavar, Kelly Berkson, Ken de Jong, Primary: Two faculty to be hired

Team 5: Research on video tagging and motion picture processing
Katy Börner (Professor of Information Science), David Crandall (Assistant Professor of Informatics), Additional: One faculty to be hired

Team 6: Development of further SLT tools
John Paolillo (Associate Professor of Informatics), Damir Cavar (Associate Scientist in Linguistics), Sandra Kuebler (Associate Professor in Linguistics), Markus Dickinson (Associate Professor in Linguistics), Russell Valentino (Professor and Chair of Slavic), Larry Moss (Professor in Mathematics), Victor Raskin (Distinguished Professor of English and Linguistics, Purdue), Julia Taylor (Assistant Professor in Computer and Information Technology, Purdue)

Team 7: Development of materials for language teaching/learning support
University Information Technology e-Learning Design Services: Umida Khikmatillaeva

Team 8: Contacts with industry
Keith Davis (Director of The Johnson Center for Entrepreneurship in Biotechnology)
Sustainability
We have discussed the project goals and the outcomes with colleagues from various academic institutions and private organizations. Due to time restrictions for this pre-proposal, we did not get endorsement letters or letters of intent from them, but we are certain that we can add these to a full GC-proposal by April next year. (…)

The researchers in the different private sector divisions expressed their interest in joint projects related to the language resources that would be created at IU (e.g. speech corpora, digital formal documentations). Given the projections of strong growth in the domain of speech and language technology, we expect the potential cooperation with the private sector to grow in the future, which will contribute to the sustainability of the envisioned language resource infrastructure, and speech and language technology research division.

Prof. Keith Davis (keirdavi@indiana.edu) from the IU’s Johnson Center for Entrepreneurship in Biotechnology expressed readiness to support and facilitate future interactions with the private sector within the scope of this project.

During 2014 and 2015, we applied for grants (…) in the domain of SLTs and language resources. The funding for research from these agencies for SLTs and language resources will likely increase in the next years. We intend to continue seeking funding for emerging research issues in the domain of the proposed project and also infrastructure development grants from these institutions.

The infrastructure and services created in cooperation with Library Technologies and ATM will continue providing the services, and there is an independent budget plan and revenue generating concept in place to guarantee sustainability of these services.

Partners
Professor Victor Raskin and Assistant Professor Julia Taylor of Purdue University will participate in this project and contribute with the expertise in Natural Language Processing and Computational Semantics. Project members have collaborated with Victor Raskin in the past on activities related to Computational Linguistics and Language Technologies. The team at Purdue is among the most prominent teams in the country for ontological semantics and computational knowledge representation, an area that is not sufficiently represented at IU.

The project creates resources and technologies that will be of great interest to various international organizations and research projects. This includes, for example, the European CLARIN project. The researchers in the current proposal are in touch and networking with the institutions and project managements of such projects, as well as with various international archives that curate and manage language data and resources.

The resources and infrastructure that will be developed at IU will also be of interest for many linguists, anthropologists, speech and language technology experts, researchers in the domain of visual image processing, cognitive computation, or digital humanities. We expect a rich partnership between individual researchers, research groups, and private sector R&D-labs to emerge during the envisioned project phase. As mentioned above, the team members of this project have experience from collaborative grants with researchers and research groups worldwide, as well as with private sector R&D divisions.
**Metrics**

The project will seek a cooperation with IU’s Center for Evaluation & Education Policy (CEEP) to develop a detailed evaluation plan and assessment metrics for all sub-projects and project outcomes.

The project generates resources that are potentially of high value to the research community worldwide, as well as to speech and language technology industry and various government agencies and divisions. It is necessary to **promote the generated resources** among those interest groups and to **evaluate the demand for data, algorithms or technology** using web-based metrics, request counts, and contacts.

The **common metrics in speech and language technology research** will be applied to evaluate technical properties of algorithms and models. These metrics include data-based comparative evaluations and general efficiency measures.

The research outcomes will be evaluated using common metrics that take into account the **number of presentations, publications, and acquired research grants**. The described project will potentially acquire substantial funding via research grants, crowd funding, or donations. The funding acquired during the initial years will be tracked and reported. The sustainable part of the project consists of services and resources that will be made available to all campus-based units and it contains potential revenue generating elements. These can be measured and used in evaluation metrics during the project. Reliable projections of these metrics can be generated for subsequent years.

Since the project is seeking collaboration and partnership with the private sector, it will track the number of collaborative projects and grants with industry partners.

Additional evaluation metrics will be worked out for all sub-projects and deliverables on a micro-level.